Causal inference study: heat pump terminology in heat-pump-home listings and sale price

Executive Summary

Using the updated heat-pump-home classification, we reran the analysis for ducted heat-pump homes with observed sale prices across all U.S. states. In this nationwide rerun, mentioning heat pumps in listing remarks is associated with a small sale-price lift after adjusting for other home and market factors. The estimated lift is roughly +0.6% to +1.0%. On a typical sale price of about $399k, that's approximately +$2.3k to +$3.9k per home, with uncertainty that could include very small negatives and modest positives (roughly -$1k to +$7k).

Because the updated heat pump flag produced a more stable treated/control structure nationwide, we did not restrict the analysis to a Northeast / Mid-Atlantic subset and did not report state-by-state causal estimates in this follow-up.

Key risks to interpretation are data quality and classification: (1) heat pump classification is imperfect (false positives/negatives); (2) sale price is inferred through deterministic linkage of listing-price fields to sold listings; and (3) duplicate or concurrent MLS listings can introduce noise. These issues are the main threats to precision and causal interpretation.

Question: For U.S. homes classified as heat pump homes that sold in 2024–2025, does including heat pump-related language in MLS remarks increase sale price?

Treatment: Binary indicator has_hp_terms (at least one heat pump term appears in listing remarks).

Outcome: Sale price (reported in dollars; modeling uses log(sale price) to estimate percent lift and then converts to dollars).

Scope change: This rerun uses the updated heat-pump-home flag and runs nationwide (all states). Unlike the initial report, we do not rely on a restricted regional subset or state-by-state effect estimates for the headline conclusion.

Headline Effect (Recommended): Nationwide, heat-pump mentions in listing remarks are associated with a small sale-price lift of about +0.6% to +1.0% (about +$2.3k to +$3.9k on a ~$399k sale), with uncertainty that could include very small negatives and modest positives (roughly -$1k to +$7k).

Primary Caveats: Heat pump classification is imperfect (false positives/negatives). Sale price is inferred via deterministic linkage of listing-price fields to sold listings; linkage errors and multi-MLS duplicates are possible. These are the main data-quality risks to causal interpretation.

Study Design

Population: Homes classified as ducted heat pump homes that sold in 2024–2025.

Unit of analysis:. A sold listing event, with deterministic handling of duplicate / overlapping listings for the same home when necessary.

Key Covariates: Home characteristics, location proxies, and market context features used as common causes (confounders) in the causal model.

Excluded features (by design): Features plausibly post-treatment, like new owner demographics or post-outcome like days on market. More granular geographic fields, like zip code, were removed when they provided limited incremental explanatory power and made balance harder.

Identification Strategy

We target the Average Treatment Effect (ATE): the average difference in sale price we would expect if, for every sold heat pump home in scope, we could switch the listing from not using heat pump terms to using heat pump terms (holding everything else fixed). Because prices are right-skewed, estimation is done on log(sale price) and then converted back to a percent lift and an implied dollar lift for interpretability.

A causal interpretation relies on a backdoor assumption: after controlling for the observed covariates (common causes), the remaining decision to include heat pump terminology is not driven by unmeasured factors that also affect sale price. This is not provable with observational data, so the study also reports balance diagnostics, overlap checks, multiple estimators, and refutation tests.

To make the estimation more reliable, the study uses (1) covariate adjustment, (2) overlap checks with optional trimming, and (3) multiple estimators with refutation tests.

Data Quality and Classification

This follow-up reruns the study after updating the underlying heat-pump-home flag (the primary fault identified in the initial report). The analysis still depends on (a) the accuracy of the updated heat pump classification and (b) the quality of the sale price field and any deterministic linkage/dedup logic.

The dataset includes both a HP-home flag (has_HP) and a text-based HP-terms flag (has_HP_terms). The cross-tab below highlights potential classification and labeling issues relevant to interpretation.

Table 1. Cross-tab: HP Flag vs. Presence of HP Terms

HP Flag	HP Terms in Remarks	Num Addresses
False	True	2.1k
True	False	256.1k
True	True	20.1k

These rows include any record where has_hp=1 or has_hp_terms=True.

Takeaway: The heat-pump flag and heat-pump terminology do not fully agree (e.g., many HP-flagged listings do not mention HP terms), indicating classification and labeling noise that can attenuate or blur the estimated effect.

Descriptive Statistics

Before any overlap trimming, listings with HP terms differ from those without terms. Notably, both the average and median sale price is higher for listings with HP terms. This motivates adjusted causal estimation rather than relying on raw differences.

Table 2. Summary by Treatment

HP Terms in Remarks	Num Addresses	Med Sale Price
False	256.1k	$375.0k
True	22.2k	$425.0k

avg_price and med_price are based on last_sale_price in the dataset.

Takeaway: Listings that mention heat pumps are higher priced even before adjustment, but this raw gap is not evidence of impact because the groups differ on observed and likely unobserved factors.

Data Cleaning

Overlap and Trimming

Before we estimate the causal effect, we do two quick checks to make sure the treated (with-terms) and control (no-terms) groups are comparable.

First, we summarize outcomes and treatment rates. The raw treated–control price difference is not causal, but it's useful context: it shows the "starting gap" before any adjustment, and it helps interpret what changes after we restrict to better-matched listings.

Second, we assess overlap and balance using a propensity score: the model's estimate of how likely a listing is to include the terms given its observed covariates. Good overlap means we can find both treated and control listings with similar propensity scores. Poor overlap means the estimate would rely on extrapolating from dissimilar homes, which is fragile and can bias results.

When overlap is weak, we apply overlap trimming: we drop listings outside the shared propensity-score region (very close to 0 or 1) and rerun the diagnostics and estimators on the remaining sample. This typically trades sample size for credibility—removing unusual cases where we can't form reasonable "like-for-like" comparisons. The **Overlap Trimming Summary** reports how much data is removed and how balance changes after trimming, and the **Treatment Share by State table** highlights where the treated share is so high or low that overlap is inherently difficult.

Table 3. Overlap Trimming Summary

Step	PS low	PS high	Kept	Dropped	Drop Rate
Overlap	3.75%	20.40%	145.4k	133.0k	47.77%

PS low and PS high are the kept propensity score bounds; Drop rate is the fraction removed by trimming.

Table 4. Treatment Share by State

State	N pre	With pre	No pre	Rate pre	N post	With post	No post	Rate post
VA	24.1k	2k	22.1k	8.36%	21.8k	1.9k	20k	8.52%
FL	43.5k	1.8k	41.7k	4.07%	23.6k	1.6k	22k	6.75%
NC	72.2k	2.4k	69.8k	3.27%	23.5k	1.5k	22k	6.47%
OH	10.8k	1.2k	9.7k	10.85%	9.9k	912	9k	9.17%
MD	19.1k	939	18.1k	4.93%	12.9k	844	12k	6.56%
CA	5.9k	856	5k	14.55%	5.5k	722	4.7k	13.21%
PA	7.9k	1.6k	6.4k	19.65%	5k	615	4.4k	12.21%
TX	4.2k	688	3.5k	16.42%	3.8k	583	3.2k	15.22%
IN	6.7k	541	6.2k	8.03%	6.6k	538	6.1k	8.17%

The report is sorted by post-trim treated count (top 15).

Takeaway: Overlap trimming is substantial in the nationwide run: 47.77% of records are removed to enforce intersection overlap, leaving 145.4k records for estimation. Treated shares by state vary, which is one driver of overlap stress in a pooled national model.

Balance Diagnostics

The diagnostic table below compares treated vs control outcomes pre and post trim. Outcome differences pre-trim are not themselves causal; they are included here for transparency about the starting imbalance and the impact of trimming.

Table 5. Outcome Diagnostics (pre vs post trim)

Stage	N	With terms	No terms	Share with terms	Med log price (no terms)	Med log price (with terms)	Med price (no terms)	Med price (with terms)
Pre	278.2k	22.2k	256.1k	7.99%	12.83	12.96	$375.0k	$425.0k
Post	145.4k	12.3k	133.1k	8.45%	12.90	12.94	$399.0k	$415.0k

Standardized mean differences (SMD) summarize covariate imbalance between treated and control. Below are the top features by absolute SMD.

Table 6. Top Covariate Imbalances (SMD)

Feature	Mean (with terms)	Mean (no terms)	SMD	Abs SMD
state_TX	0.0475	0.0244	0.124	0.124
state_NC	0.1238	0.1653	-0.118	0.118
climate_zone_Other	0.0756	0.0474	0.118	0.118
state_CA	0.0588	0.0356	0.109	0.109
state_FL	0.1297	0.1652	-0.100	0.100
n_units_nan	0.0554	0.0356	0.095	0.095
state_KY	0.0164	0.0307	-0.095	0.095
state_CO	0.0169	0.0071	0.089	0.089
home_type_TH	0.0459	0.0660	-0.088	0.088
state_PA	0.0501	0.0332	0.084	0.084

Takeaway: After overlap restriction, treated share rises from **7.99% to 8.45%** and outcome medians remain separated, **$415k vs $399k**; several covariates still show moderate imbalance, motivating emphasis on the more robust estimators.

Estimators

All estimators report an ATE on log(sale price). Interpreting effects as percent lift: if the ATE is 0.018 on log(price), that corresponds to roughly a 1.8% increase in price. We then translate percent lift into dollars by applying it to a baseline sale price (here, the post-trim median price of the control group).

Double Machine Learning (DML) is designed for settings with many covariates and flexible nuisance models. It first uses machine learning to predict (a) treatment propensity and (b) expected outcome, then "orthogonalizes" the final effect estimate so that small errors in those nuisance models have less impact on the estimated treatment effect. This often improves robustness relative to a single regression specification when the true relationships are non-linear or high-dimensional.

Doubly robust (DR) learners combine an outcome model and a propensity model; the effect estimate remains consistent if at least one of those nuisance models is correctly specified (under standard regularity conditions). In practice, this provides a useful cross-check against relying on a single modeling approach.

All estimators are implemented with backdoor adjustment. The study reports:

- Baseline regression: linear adjustment using the specified common causes.
- LinearDML: orthogonalized estimation that uses ML for nuisance models and targets a linear treatment effect.
- DRLearner: doubly robust learner combining outcome and propensity models; consistent if either nuisance model is correct.

Nuisance models for outcome and propensity are fit with an HGB (hist-gradient-boosting) family.

Causal Effect Estimates

Effects are estimated on log(sale price) and then converted to a percent lift for readability. For small effects, an ATE of 0.029 on log(price) is approximately a 2.9% increase in price. We translate percent lift into dollars by applying it to a baseline sale price; throughout, we use the post-trim median sale price of the control group, $399k.

Table 7. ATE Estimates

Estimator	ATE (% lift)	CI 95% Low	CI 95% High	$ Lift	$ Lift Low	$ Lift High
Baseline	0.25%	-0.77%	1.28%	$998.30	$-3.1k	$5.1k
Linear DML	0.57%	-0.29%	1.44%	$2.3k	$-1.1k	$5.7k
DR Learner	0.99%	0.15%	1.83%	$3.9k	$609.60	$7.3k

Across the ML-assisted estimators (LinearDML and DR Learner), the effects are small. LinearDML's 95% CI includes 0%, while DR Learner's 95% CI is entirely positive. Given the small magnitudes, remaining imbalance, and sensitivity to estimator choice, the nationwide result does not support a strong causal claim that mentioning heat pumps in remarks increases sale price in a systematic way in the overlap-supported sample.

Takeaway: We treat the ML-assisted estimates as the primary evidence. Point estimates range from +0.57% to +0.99% (about +$2.3k to +$3.9k on a $399k median). LinearDML's 95% CI crosses 0%, while DR Learner's 95% CI is entirely positive (0.15% to 1.83%).

Refutation Tests

Refuters are not definitive proofs, but they help detect obvious failure modes. The placebo treatment refuter should produce an effect near 0; the other refuters assess stability to resampling or adding random confounding.

Table 8. Refuter Summary

Estimator	Placebo New Effect	Placebo p	Random Common Cause New Effect	RCC p	Subset New Effect	Subset p
Baseline	-0.7%	0.0	0.2%	0.0	0.1%	0.0
Linear DML	-0.7%	0.0	0.6%	0.0	0.2%	0.0
DR Learner	-0.7%	0.0	0.9%	0.0	0.4%	0.0

The pooled refuters show stable results. The placebo test pushes the effect toward ~0, showing that the pipeline isn't "finding an effect no matter what." The other refuters (random common cause, subset) keep the effect positive and in the same general neighborhood, while still showing that the exact magnitude can move around under perturbations.

Takeaway: pooled refuters are directionally reassuring: *placebo collapses the effect; non-placebo stress tests keep a positive effect, with moderate sensitivity in size.*

Limitations and Risks

- Heat pump classification noise. Some homes labeled as heat pump homes may not truly have heat pumps; conversely, some homes may mention

heat pumps in remarks without being detected in structured features. This can attenuate or distort estimated effects.
- Sale price linkage. We do not always have a clean single final sale price per listing episode; we link the last observed listing price to the sold listing. Linkage errors can introduce measurement error.
- Duplicate/concurrent listings. The same home can appear across MLS feeds simultaneously. We deterministically select one listing record, which may not match the record buyers actually saw.
- Unobserved confounding. Even with extensive covariates and trimming, we cannot rule out omitted variables correlated with both wording choice and price (e.g., agent sophistication, listing quality, renovations not captured in structured data).
- Generalization. Estimates apply to the overlap-supported subset after trimming and to the restricted geography; effects may differ in other regions.

Recommended Next Steps

- Validate heat pump classification against higher-precision signals (permits, HVAC service records, equipment registries) for a subset to quantify misclassification rates.
- Validate sale-price linkage and duplicate-listing selection with manual spot checks in states with the largest estimated effects and/or highest listing duplication.
- Explore richer text features (term families, counts, or a classifier-based 'heat pump description quality' score) and heterogeneity (by state, price tier, home type).
- If feasible, run a forward test with a partner where heat pump language templates are randomly encouraged or policy-driven to reduce remaining confounding.

Technical Appendix

Heat Pump Terms

- heat pump
- heat pumps
- hpwh
- heat pump water heater

- air source heat pump
- ground source heat pump
- geothermal heat pump
- variable refrigerant flow
- vrf
- inverter heat pump
- cold climate heat pump
- hybrid heat pump
- high efficiency hvac
- electric heat pump
- heat pump dryer
- daikin fit
- mitsubishi hyper heat
- bosch ids
- electric furnace

Causal & Modeling Python Packages

- **DoWhy**
  - Used to define the causal graph (treatment/outcome/common causes), run a baseline backdoor estimator, and run refutation tests.
- **EconML**
  - Used for the primary causal estimators: **LinearDML**, **DRLearner**, and **LinearDRLearner** (ATE on log-price with conversion to %/$ for reporting).
- **scikit-learn**
  - Used for nuisance modeling (propensity and outcome models), and general model utilities.

Causal Estimators

- Baseline: backdoor.linear_regression (DoWhy)
- Primary:
  - econml.dml.LinearDML
  - econml.dr.DRLearner
  - econml.dr.LinearDRLearner
  - econml.dml.CausalForestDML

Overlap Trimming

- Propensity-family: overlap_family="hgb"
- Cutoff: overlap_cutoff=99
- Mode: overlap_trim_mode="intersection" (keeps the shared propensity region between treated/control)

Key Scikit-Learn Hyperparameters

Used both for overlap trimming and as the propensity component inside the causal estimators (binary treatment):

- HistGradientBoostingClassifier:
  - max_depth=3
  - min_samples_leaf=200
  - max_iter=300
  - learning_rate=0.05
  - early_stopping=True
  - validation_fraction=0.1
  - n_iter_no_change=20

Refuters (DoWhy)

- Placebo treatment refuter (placebo_type="permute")
- Data subset refuter (subset_fraction=0.5)
- Random common cause refuter
- Defaults:
  - num_simulations=20

Reproducibility

- Global random seed used throughout: random_state=257